

In Silico Evolution of Early Metabolism

Alexander Ullrich**

University of Leipzig

Markus Rohrschneider†

University of Leipzig

Gerik Scheuermann†

University of Leipzig

Peter F. Stadler*,*,‡,§

Max Planck Institute for Mathematics
in the Sciences

Christoph Flamm§

University of Vienna

Abstract We developed a simulation tool for investigating the evolution of early metabolism, allowing us to speculate on the formation of metabolic pathways from catalyzed chemical reactions and on the development of their characteristic properties. Our model consists of a protocellular entity with a simple RNA-based genetic system and an evolving metabolism of catalytically active ribozymes that manipulate a rich underlying chemistry. Ensuring an almost open-ended and fairly realistic simulation is crucial for understanding the first steps in metabolic evolution. We show here how our simulation tool can be helpful in arguing for or against hypotheses on the evolution of metabolic pathways. We demonstrate that seemingly mutually exclusive hypotheses may well be compatible when we take into account that different processes dominate different phases in the evolution of a metabolic system. Our results suggest that forward evolution shapes the metabolic networks in the very early steps of evolution. In later and more complex stages, enzyme recruitment supersedes forward evolution, keeping a core set of pathways from the early phase.

Keywords

Metabolic evolution, genotype-phenotype map, multi-scale model, evolution of catalysis

A version of this paper with color figures is available online at http://dx.doi.org/10.1162/artl_a_00021. Subscription required.

1 Introduction

Understanding the evolutionary mechanisms of complex biological systems is an intriguing and important task of current research in biology as well as artificial life. The mechanisms that governed the formation of metabolic pathways from chemical reactions have been discussed for decades, and several hypotheses have been proposed since the 1940s. Research on the TIM β/α -barrel fold architecture [6], for instance, shows that the evolution of modern metabolism is mainly driven by enzyme recruitment, as suggested by the patchwork model [31, 55]). Gene duplications, on the other hand, may facilitate the specialization of an originally multifunctional enzyme, such as carbamoyl phosphate

* Contact author.

** Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. E-mail: alexander@bioinf.uni-leipzig.de (A.U.); studla@bioinf.uni-leipzig.de (P.F.S.)

† Image and Signal Processing Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, PF 100920 D-04009 Leipzig, Germany. E-mail: rohrsneider@informatik.uni-leipzig.de (M.R.); scheuermann@informatik.uni-leipzig.de (G.S.)

‡ Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. E-mail: stadler@mis.mpg.de. Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany. Center for Non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark. The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501.

§ Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria. E-mail: studla@tbi.univie.ac.at (P.F.S.); xtof@univie.ac.at (C.F.)

synthetase, to diverse functions in new pathways [27]. Similarly, an entire metabolic pathway may duplicate and specialize, as has been shown to be the case for tryptophan and histidine biosynthesis [18, 31]. The ability of enzymes to catalyze reactions other than those for which they are physiologically specialized, dubbed *enzyme promiscuity* [33], forms an important evolutionary reservoir from which novel catalytic functions can be drawn. Promiscuous enzyme activities, although far less efficient than well-evolved ones, can be assembled into novel metabolic pathways [34], which can provide a selective advantage in particular environmental niches. The evolutionary potential of enzyme promiscuity thus extends far beyond mere enzyme recruitment. Based on such evidence from modern metabolic networks, several hypotheses to explain the evolution of metabolism in general and the emergence of specific metabolic pathways have been suggested. The four most widely cited scenarios are briefly discussed in the following paragraphs. For an in-depth discussion of these theories and further examples of pathway evolution we refer the interested reader to three recent review articles [4, 9, 44].

The *backward evolution hypothesis* was one of the first theories of the evolution of metabolic pathways, proposed by Horowitz [26]. It assumes that an organism is able to make use of certain molecules from the environment. However, individuals that can produce these beneficial molecules by themselves gain an advantage in selection in the case of depletion of the food source. Therefore, new chemical reactions are added that produce beneficial molecules from precursors that are abundant in the environment or that are produced in turn by the organism's metabolism. As a consequence, one should observe more ancient enzymes downstream in present-day metabolic pathways. Toward the entry point of the pathway, younger and younger enzymes should be found (see Figure 1a). Backward evolution has been proposed for both the glycolytic pathway [15] and the mandelate pathway [39].

The *forward evolution hypothesis* can be seen as an extension or counterpart of the backward evolution hypothesis, reversing the direction of pathway evolution. Granick [19], and later Cordon [8], argued for a pathway evolution in the forward direction, requiring that the intermediates be already beneficial to the organism. This is in particular plausible for catabolic pathways, where the organism can extract more energy by breaking food molecules down to simpler and simpler end products. Older enzymes are then expected to be upstream in the pathway, with younger enzymes appearing further downstream (see Figure 1b). The isoprene lipid pathway [38] is an example for the development of biosynthetic pathways in the forward direction.

The *patchwork model* [31, 55] explains the formation of pathways by recruiting enzymes from existing pathways. The recruited enzymes may change their reaction chemistry and metabolic function in the new pathways and specialize later through evolution. This introduction of new catalytic activities leads to a selective advantage. Looking at the constitution of a pathway formed by enzyme recruitment, we should observe a mosaic-like picture of older and younger enzymes mixed throughout the pathway (see Figure 1c). The observation that the TIM β/α -barrel fold architecture occurs in many different pathways corroborates widespread enzyme recruitment in modern metabolism [4]. Other examples are pyrimidine metabolism and histidine biosynthesis [39].

The *shell hypothesis* was proposed by Morowitz [35]. It argues, for the case of the reductive citric acid cycle, that in the beginning an autocatalytic core is formed from which new catalytic activities and pathways can be recruited and fed. Thus a metabolic shell would form around this core. Enzymes in the core would likely be less prone to mutational changes, because they are essential for the organism. Thus, one should still be able to observe a core of ancient enzymes (see Figure 1d). According to Morowitz [35], the reductive citric acid cycle constitutes such an autotrophic synthetic system.

Each of these mechanisms, thus, is supported by evidence for certain pathways, so that none of these mechanisms is exclusive. Studies on hypotheses of pathway evolution [4, 35] suggest that metabolism has evolved differently in different phases. Furthermore, only traces, or "shadows," are still observable from the events in the very distant past of terrestrial life. Many aspects of the evolutionary history are therefore still not well understood. In particular, the first steps that lead to the emergence of the earliest forms of metabolism evade observation by conventional approaches.

Thus, there is an urgent need for detailed and realistic models of early metabolism that consider all its components and scales. Simulation approaches have been shown to be useful in finding and challenging explanations for the evolution of biological networks [40]. We have recently

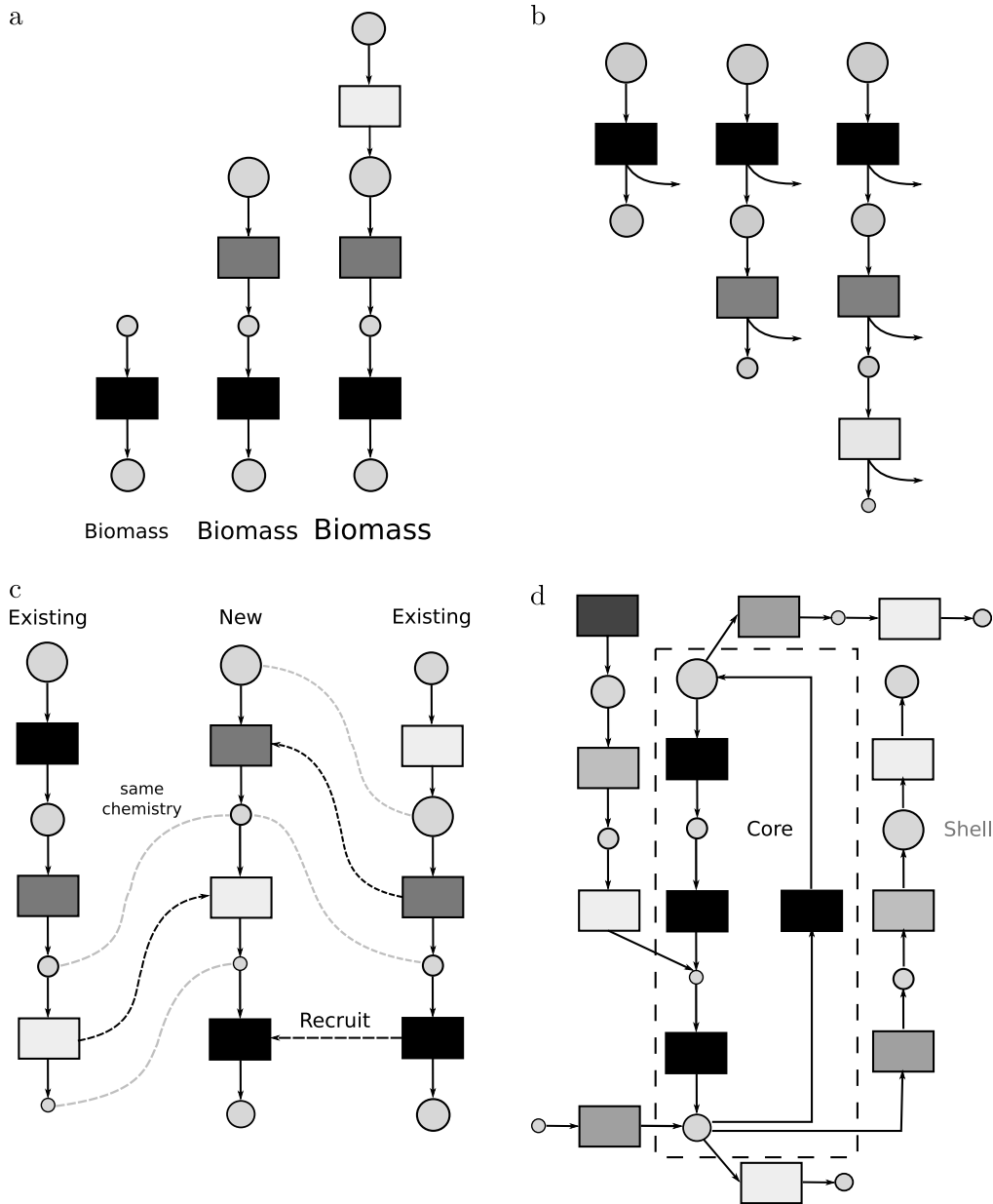


Figure 1. Hypotheses about the formation and evolution of metabolic pathways: (a) backward evolution, (b) forward evolution, (c) patchwork model, (d) shell hypothesis. Squares represent enzymes; gray circles denote metabolites. Grayscale encoding for enzymes stands for their age, dark being older and light being younger.

proposed a computational framework for the evolution of metabolism [12], modeling all its significant components in a realistic way. Here, we focus on the detailed analysis of evolutionary transitions, aiming in particular at an understanding of the processes underlying metabolic innovation.

2 Model

Innovation is hard to model. In contrast to population dynamics or quantitative genetics, where the dynamics is governed by Darwinian selection and the generation of variability can be described by

simple statistical models, we need a way of judging whether an innovation has been selected, or whether an observed fitness increase is the result of an incremental adaptation. This implies that phenotypes must be represented explicitly as objects whose fitness can be evaluated. This paradigm has been explored two decades ago in the context of evolving RNA molecules [13], where phenotypes are modeled as RNA secondary structures. Subsequent investigations have demonstrated that the sequence-structure map or, in a more general setting, the genotype-phenotype map [29, 46] plays a crucial role. More recently, neutral networks were studied in the context of gene regulation [5] and metabolic networks [43].

In particular, the accessibility of potential novelties plays a crucial role: In realistic settings the search spaces are so large that evolutionary trajectories are determined to a large extent by the ease with which advantageous mutants can be generated from extant populations [14, 48]. It is crucial, therefore, to devise models of phenotypes that are as biophysically realistic as possible and computationally feasible. While in the case of RNAs the fairly simple and well-understood relation of RNA sequences and RNA secondary structures (i.e., the map defined by RNA folding) could be employed, it is a highly complex task to devise realistic genotype→phenotype→fitness mappings for even minimal organisms. A very primitive *ribo-organism* was devised, for instance, to study the evolution of primitive genetic codes [54].

Clearly, a completely realistic quantitative simulation of all aspects of the phenotype is infeasible. Even RNA secondary structures are only rough caricatures of real, three-dimensional molecules in all their quantum-chemical glory. Detailed realism, however, is not required for our purposes: It suffices that the probabilities of accessibility, the variabilities among adjacent phenotypes, and similar statistical properties of genotype-phenotype maps are modeled in a reasonably realistic way [50].

In order to study the origins of a metabolism (our problem at hand), several distinct levels of description need to be modeled: organisms that are subject to selection, genomes that account for genetic variability and encode catalytic enzymes, and a system of chemical reactions, some of which are catalyzed by the enzymes. The computational model, summarized schematically in Figure 2, is primarily composed of a genetic and a metabolic subsystem [12].

The genetic subsystem is implemented as a cyclic RNA genome. A special sequence motif indicates the start of genes, which are of constant length. The RNA sequence corresponding to the coding sequence of a gene is folded into the (secondary) structure using the Vienna RNA Package [25] (step A in Figure 2).

The RNA genes encode for ribozymes. (For simplicity, we use a *ribo-organism* model here to avoid the additional complications of a genetic code, a translation system, and the need to model some form of protein folding.) The ribozymes act as catalysts in the chemical function of the metabolic subsystem. A major problem, for which no satisfactory physical theory is available, is the relationship of the 3D structure of an RNA (or protein) and the catalytic function that it exerts. Some observations from evolutionary studies of catalytic molecules, however, suggest a simplified heuristic that can be used to connect the specific structural features of the ribozyme with its catalytic activity in a way that reproduces crucial statistical features of reality [52].

A chemical reaction can be characterized by an intermediary transition state structure graph that describes the atom and chemical bonds involved. During chemical reactions bond formation/breaking is confined to a small subset of atoms of the reacting molecules. A cyclic graph abstraction, called the imaginary transition state (ITS) [16], can be used to capture the changes in the reactive center [21]. Furthermore, over 90% of all known organic reactions can be classified by their ITS [22] and organized in a hierarchical structure [23]. Enzymatic functions, on the other hand, are largely determined by an active center and its molecular properties. The active center typically comprises only a small part of the ribozyme—in our model, the largest loop of the RNA secondary structure. Furthermore, most ribozymes as well as protein enzymes act by binding and stabilizing the transition state [53], thereby decreasing the activation energy of the reaction. Hence we relate the longest loop of the RNA secondary structure to the imaginary transition state. To this end, sequence and structure features of the folded RNA gene products are mapped into the classification tree of organic reactions for functional assignment of the catalytic set (step B in Figure 2).

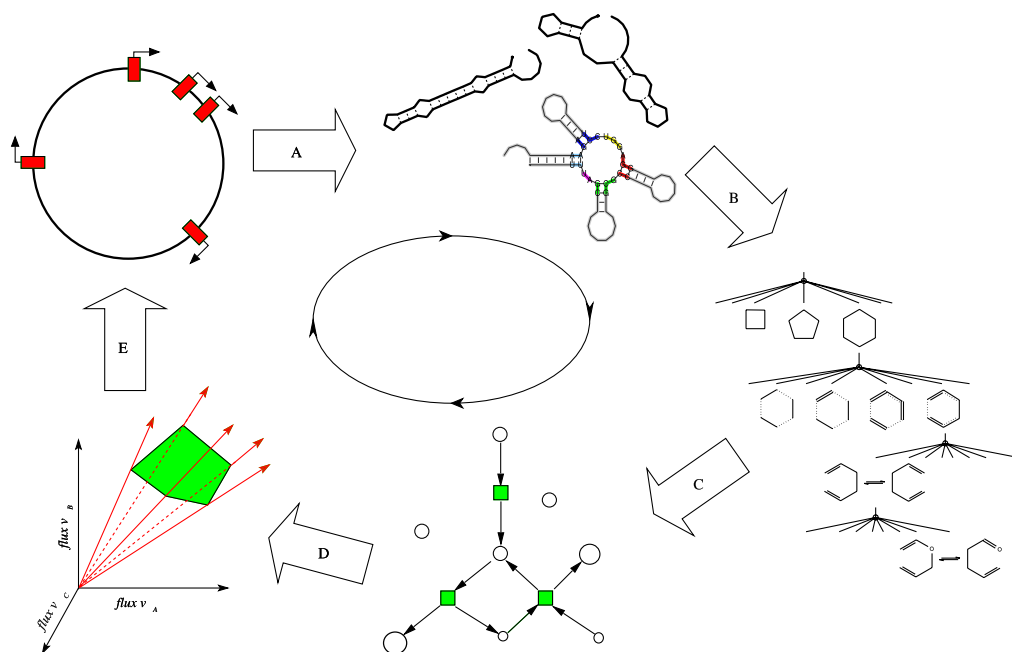


Figure 2. Scheme of the simulation system. A: decoding of (RNA) genes to catalytic molecules; B: assignment of catalytic functions to ribozymes, through mapping from structural and sequential information of the RNA molecule to a reaction logo in the hierarchy [21]; C: construction and stochastic simulation of the metabolic network; D: metabolic flux analysis and fitness evaluation; E: application of genetic variation operators.

The structure-function map is hence implemented by treating the ITS tree as a kind of decision tree. Its branches are chosen based on specific structural information of the loop region. The decision process terminates in a leaf that corresponds to the transition state of the catalyzed reaction. The ITS hierarchy consists of four levels. The first level is the size of the transition state, which indicates the number of molecules that are actually involved in the electron reordering during the reaction. The length of the loop region determines the size of the transition state. The number of adjacent stems in the loop region and specific enclosing base pairs encode for the bond types of the transition state (level 2). Further, the positions of these stems in the loop specify the position of the bonds (level 3). For the fourth level, the sequence inside the loop is analyzed to assign the atom types of the transition state.

The relation between genotype and phenotype is determined by RNA folding. This RNA sequence-to-structure map has been studied extensively [41, 45, 46] and exhibits a high degree of neutrality, extensive neutral networks, and shape space covering. Hence almost all possible secondary structures can be reached through neutral evolution, that is, changes in the RNA sequence without loss of function. Furthermore, the neutral networks of any two structures nearly touch somewhere in sequence spaces, while distances can be large in other regions. This structure governs the dynamics of evolution and to a large extent dominates the superimposed structure-to-function map [12, 49]. Overall, the composition of the RNA-sequence-to-structure map and the structure-to-function map detailed above is consistent at least qualitatively with the observed sequence dependence of enzyme catalysis [52].

It is important to note that this implementation of the relation between sequence and catalytic function is evolvable in several ways: Mutations can affect both substrate specificity and the type of the chemical reaction itself, thus allowing true innovations of novel catalytic functions. In particular, they allow the metabolic organization to escape from the confines of the chemical space set by the initial conditions of the simulation.

The metabolic subsystem is built upon a graph-based artificial chemistry [2] endowed with a built-in thermodynamics. To generate the metabolic reaction network, induced by the catalytic set (chemical reactions decoded from the genome) on the set of metabolites (chemical molecules of interest from user input), a rule-based stochastic simulation is performed, where the likelihood of a reaction being chosen depends on its reaction rate [10]. Reaction rates are calculated “on the fly” from the chemical graphs of the reactants (step C in Figure 2).

Fitness evaluation, finally, is based on flux balance analysis [11, 37]. To identify the elementary flux modes—that is, the extreme pathways [17]—of the resulting reaction network, a metabolic flux analysis is performed (step D in Figure 2). The fitness of an organism is computed as the maximum of the (linear) yield function (e.g., biomass production) over all extreme pathways. Finally, genetic variation operators (i.e., mutation and—in some simulation scenarios—recombination) are applied to the genome (step E in Figure 2). Variants are then selected depending on their fitness.

The computational model outlined here, in summary, implements a complex fitness function that deterministically depends on the genomic sequence and on the environment. The latter is determined by the chemical composition of the food source on which our simplified organisms live. Despite their complex internal structure, hence, our model organisms still show no complex behavior such as predation or any form of regulated adjustment to their ecosystem—they simply compete with a fixed strategy for the common, albeit time-variable, chemical resources.

3 Simulations and Results

In this section we use the computational model described above to simulate the evolution of metabolic networks and analyze the change of its structure and components over several generations. All simulation runs performed for this article were initialized with the full set of chemical reactions to choose from, the same configurations for genome length (5000 bases), and the same TATA-box constitution (UAUA) and gene length (100 bases). They differ in initial conditions, population size, environmental conditions, selection criteria, and simulation time (number of generations).

3.1 Quantitative Analysis

To gain some quantitative insights into the general principles of metabolic evolution, we performed a series of simulation runs to investigate certain measures that give a picture of the evolutionary constitution of the metabolic networks throughout the evolution process.

In a previous study [51], we already showed that our metabolic networks evolved certain properties such as a scale-free node degree distribution and the existence of hub metabolites. An investigation of the enzyme connectivity suggested that enzymes from early stages show a higher connectivity than those from later stages. Here, we confirm these findings with a much larger sample of 100 simulation runs starting from the same set of initial metabolites (cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene). Figure 3a shows a clear trend for enzymes from the first generations to be responsible for the major part of connections in the metabolic network. On the one hand, this can be explained as simply due to the fact that enzymes that enter the system earlier have more time to form connections. On the other hand, this observation could also indicate that enzymes with higher and higher specificity evolve in the later stages. It could be anticipated that enzymes with all specificities still appear in later generations but only specific enzymes catalyzing few reactions are taken to the next generation, while multifunctional enzymes are discarded because they would change the structure of the network too drastically. Considering the connectivities of metabolites (see Figure 3b), we still find the highly connected nodes in the early steps, especially if we consider environment metabolites, which are always abundant and which account for about 50% of connectivity. However, there is constant production of metabolites, potentially becoming highly connected.

In order to find arguments for some of the evolution hypotheses, we study the occurrence time (age) of reactions and metabolites along pathways. It is of particular interest to determine in which

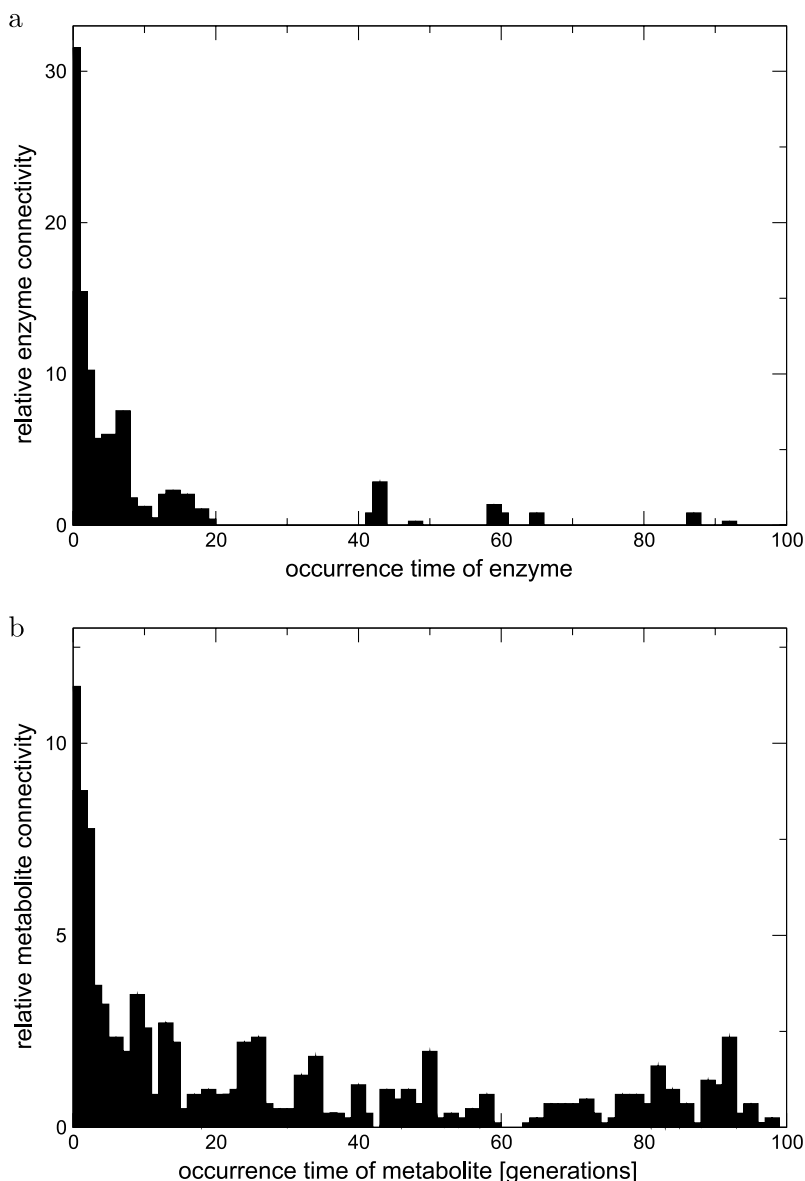


Figure 3. Average relative connectivity of (a) enzymes and (b) metabolites introduced in the same generation, for 100 generations. The height of the bars shows the fraction of the overall connections that are accounted for by enzymes/metabolites from a particular generation. All values are averages over 100 simulation runs. Input molecules are not considered in the statistic; they account for nearly 50% of metabolite connectivity.

direction—downward (with the flow of mass), or upward (against the flow)—pathways are formed by addition of chemical reactions that recruit or produce new metabolites. We will use the term forward (backward) link if, in a pair of reactions in a pathway, the successor is evolutionarily older (younger). In the same vein, a forward (backward) link between metabolites refers to a situation in which the products of a reaction are evolutionarily older (younger) than the educts. Accordingly, we define forward (backward) pathways as pathways in which there is at least one forward (backward) link and no backward (forward) link. Given these definitions, we compute the set of extreme pathways for every generation and all cells. For each pathway we then determine the percentage of forward and backward links and pathways, for both reactions and metabolites.

For this study, we performed 100 runs with the following settings: The population was 100 cells running for 100 generations and performing 100 network expansion (stochastic simulation) steps per generation; the input molecules were cyclobutadiene, ethenol, phthalic anhydride, methylbutadiene, and cyclohexa-1,3-diene. In Figure 4 we see the change from generation to generation in the

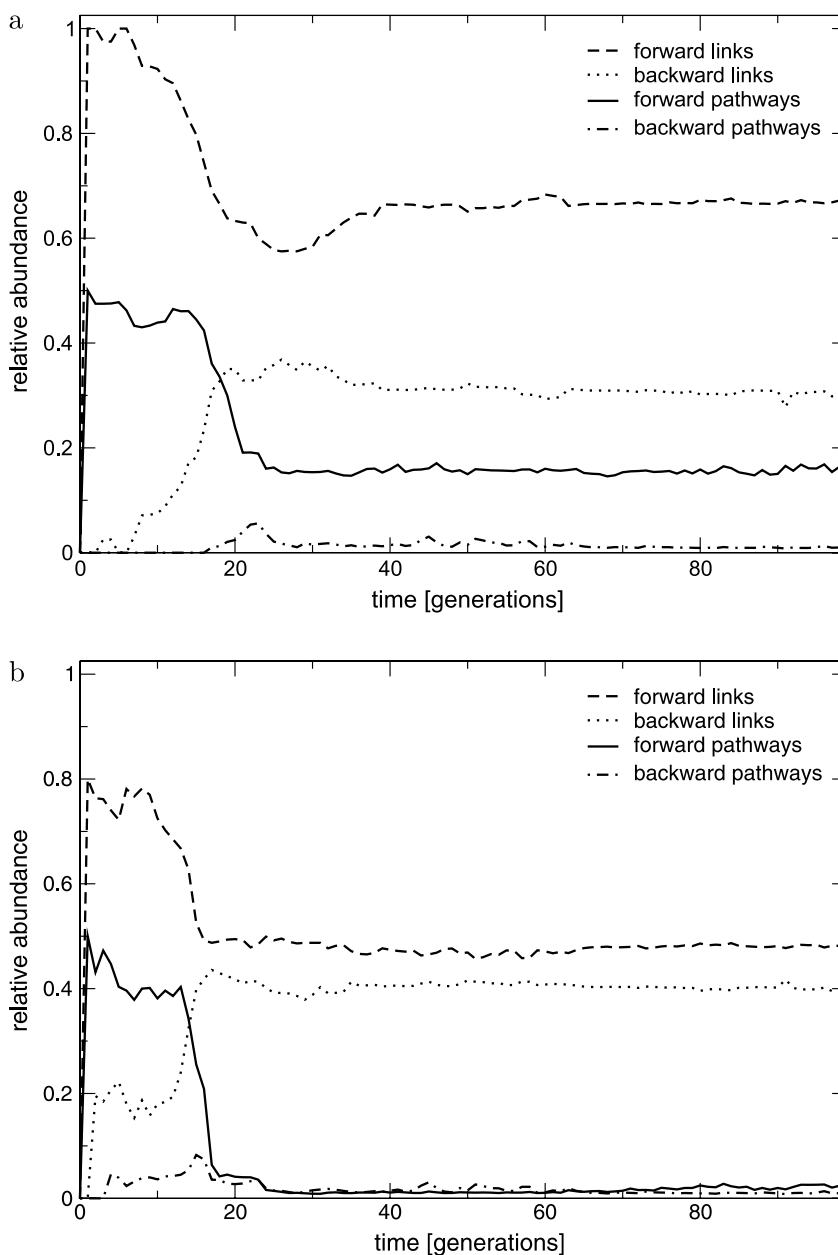


Figure 4. Evolutionary history of simulated metabolic networks. For the first 100 generations, we show the numbers of links and pathways that conform to the forward and backward evolution scenarios, respectively. Links are pairs of (a) consecutive reactions or (b) consecutive metabolites along a pathway. A pathway is identified as *forward-evolved* if at least one of its links is forward and none backward. In the first generations, the network consists predominantly of forward (reaction) links and pathways. After about 20 generations, the relative abundance of forward pathways decreases drastically but quickly reaches a persistent plateau value.

constitution of the metabolic networks regarding our measures of forward and backward links and pathways. Considering the reactions of the networks, we observe that in the first generations, the networks consist mainly of links and pathways conforming to the forward evolution scenario. However, in later generations we observe a much more mixed mosaic-like picture arguing in favor of the patchwork model. This trend becomes even more evident from the metabolite's point of view: Almost all pathways consist of forward and backward links in equal numbers. Another observation from the reaction's point of view is that most forward pathways from the early stages remain even in the last stages, which could mean that they form a core of pathways that are not subject to evolutionary change. This supports the shell hypothesis.

For the gene history analysis in the next section we also performed longer simulation runs of 2000 generations with the same initial conditions as in the previous simulations, but with fewer network expansion steps per generation. The statistics for these runs are summarized in Figure 5. We observe the ongoing trend of a mix of forward and backward links while retaining a certain fraction of forward pathways throughout evolution, giving further support to the previous observations. While the first generations of network evolution are dominated by the forward evolution scenario, patchwork evolution takes over after a sufficiently diverse repertoire of enzymes has built up from which enzymes can be recruited. An evolution in layers, as proposed in the shell hypothesis, so far has not been observed. However, the existence of the set of pathways that originated by forward evolution in the earliest generations at least suggests the possibility of an ancient metabolic core from which later pathways are built by enzyme recruitment or other strategies, such as enzyme or pathway duplication.

So far, our simulation results do not provide any support for the backward evolution scenario. However, we have not yet simulated an environment with temporary depletion of food metabolites, which is one of the major assumptions of this theory. To this end we perform a study investigating the impact of variations in resource abundance on metabolic evolution. For this, we analyze 100 simulations over 900 generations. The initial conditions are the same as in the previously described simulations. However, the set of food metabolites is changed for certain time periods. For the first 100 generations we leave the system unperturbed and thus under the same conditions as in the previous simulations. Starting from generation 100, one of the five food metabolites will be removed from the input set; this means that it can still be produced by other metabolites but is not permanently imported from the environment and might deplete like other internal metabolites. After a time interval of 50 generations, the removed metabolite is added back to the set of food metabolites, and the next food metabolite is removed. After every food metabolite has been removed once (after 350 generations), the initial food set is reintroduced for the next 50 generations. The upper plot in Figure 6 shows the metabolic pathway evolution statistics for these first 400 generations. In

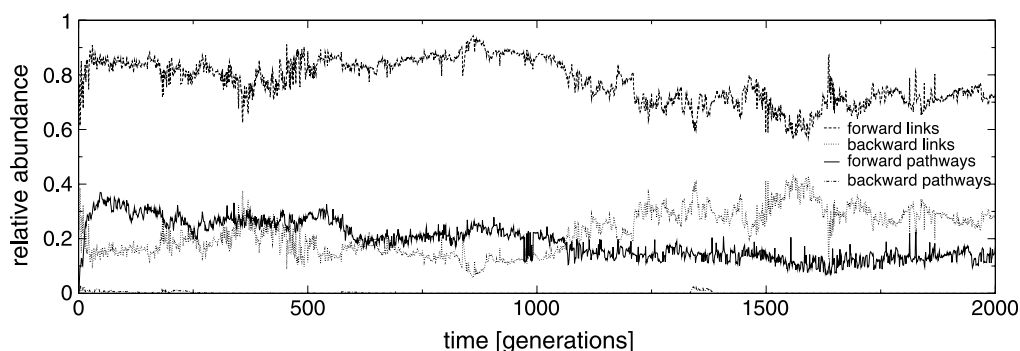


Figure 5. Evolutionary history of longer (2000 generations) metabolic network simulations. The statistics, as explained above, show the numbers of links and pathways that conform to the forward and backward evolution scenarios, respectively. The trend is similar to the shorter simulations in that forward evolution dominates the starting phase, converging to a mixed situation typical of the patchwork model.

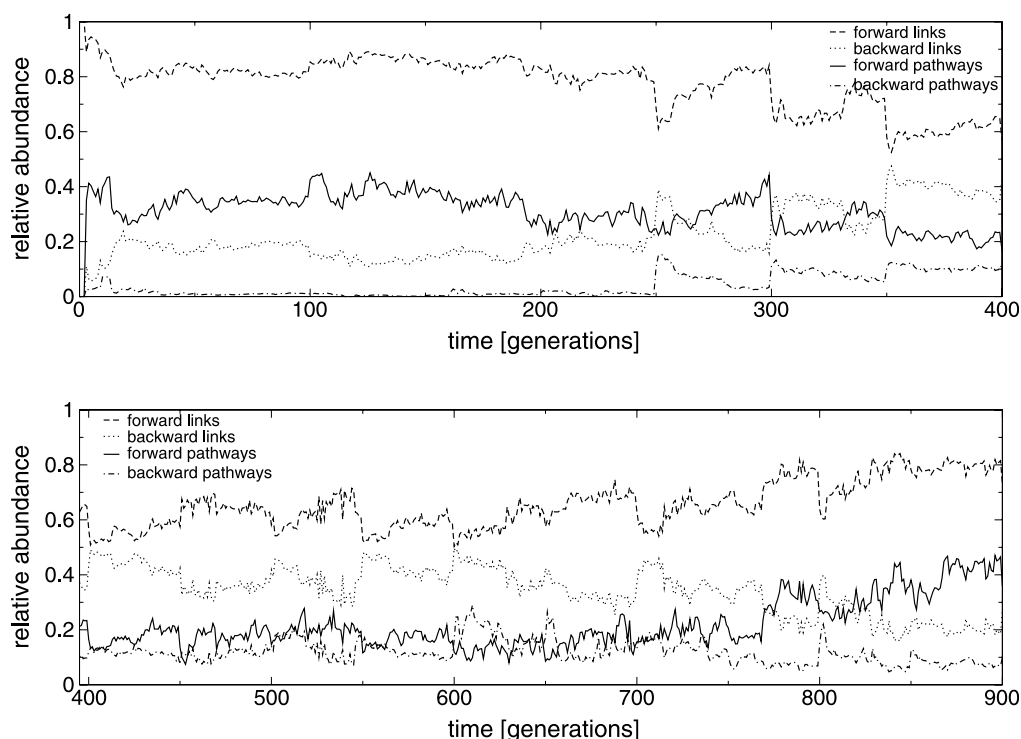


Figure 6. Simulated evolutionary history of networks in an environment with perturbations in the set of food metabolites. The same statistics as for the other simulations are used, namely, the numbers of links and pathways that conform to the forward and backward evolution scenarios, respectively. For the first 100 generations, the food set is the same as for the previous simulations. In the next five intervals of 50 generations each, the food metabolite set is perturbed by removing one of the five initial metabolites at a time. For generations 350–400, the original full set is reintroduced. The remaining simulation (400–900) is divided in intervals of 50 generations. In each interval, exactly two of the food metabolites are withdrawn. The system reacts to the largest perturbations with a sudden increase in backward links and pathways, followed by a slower decrease, suggesting that metabolite depletion can indeed induce backward evolution.

the next phase, which is depicted in the lower plot in Figure 6, pairs of food metabolites are removed for intervals of 50 generations in the same procedure as described above for the removal of single metabolites. After 900 generations every possible combination of two metabolites from the full set of food metabolites had been removed once.

In the first 100 generations, no fundamental differences from the previous simulations can be observed. This is to be expected, since the conditions up to this point are the same. In the following three intervals of 50 generations, from generation 100 to 250, still no significant changes in the network evolution are apparent, despite the perturbation in the set of food metabolites. In the following 100 generations (250–350), however, we observe an increase in backward links and the emergence of backward pathways. More specifically, at the beginning of each of the two intervals there is a sudden and drastic increase of backward-evolved pathways followed by a slow decline for the rest of the interval. The explanation for the increase in backward pathways is that of the retrograde evolution hypothesis: Due to the depletion of an important food metabolite, there is a strong selective advantage for pathways that can produce this metabolite from other sources. The sudden increase suggests that the reactions, or at least the enzymes, of these backward pathways must have been already present in the network rather than invented through the emergence of entirely new enzymes or catalytic mechanisms, that is, capitalizing on enzyme promiscuity. The other factor explaining the sharp increase in the beginning and the slow decline afterward is that the network complexity and thus the overall number of metabolic pathways decreases after depletion of the food metabolite.

Therefore, pathways that depend upon the depleted metabolite, mostly forward pathways, disappear, since they no longer contribute significantly to the organism's fitness. When the network later evolves new (forward or mixed) pathways based on the new conditions, the fraction of backward pathways among all pathways decreases.

The five intervals of perturbation are followed by an interval (350–400) with the full food metabolite set. Surprisingly, this interval also starts with a sudden increase in backward links as well as a slightly smaller increase in backward pathways and is not followed by a decrease later in the interval. We have no satisfactory explanation for this observation so far.

In the second phase (400–900), in which pairs of food metabolites are withheld, similar observations to those in the first phase of perturbations can be made. However, the modifications to the network are less clearly significant and less clearly visible. The depletion of food metabolites is again mostly followed by an increase in backward links and pathways, but less dramatic than before. Further, the slow decrease after the sharp rise is not as clear as in the previous phase and sometimes interrupted by smaller increases. In the later intervals (750–900), one can even observe an overall decrease in backward links and pathways compared to forward pathways. We conclude that synchronous depletion of multiple metabolites can induce some form of backward evolution. At the same time, however, too invasive perturbation can disrupt the entire system. Furthermore, we suggest that the backward evolution we observe in our simulations is driven by an enzyme recruitment process rather than the formation of completely novel reaction pathways.

3.2 Detailed Analysis of Simulated Evolutionary Histories

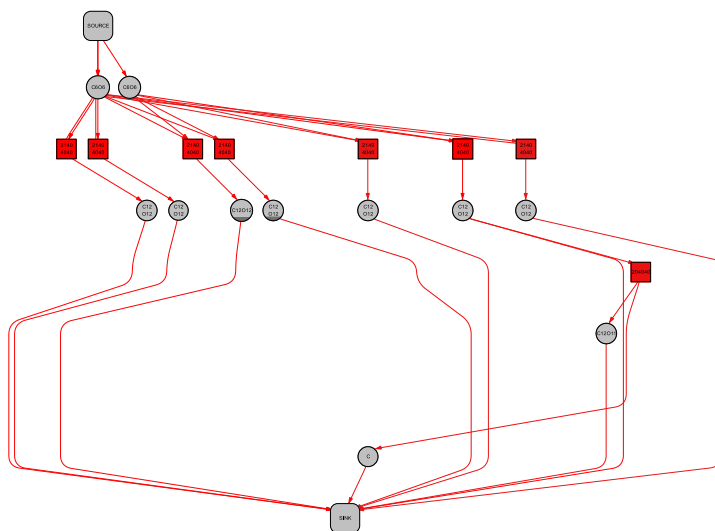
In the following we illustrate some of our findings from the previous study in more detail for a single simulation, starting with only two input molecules and developing only a few enzymes. For the visualization of an evolutionary time series see Figure 7. An animation of the network evolution is provided as supplemental material (see Section 5). Figure 8 gives an overview of reaction and metabolite lifetimes. The genome, and hence the set of enzymes, is chosen at random in the beginning. The two input molecules of this simulation are cyclic and sequential glucose. The simulation run is terminated after 100 generations.

We focus again on the evolutionary constitution of the metabolic network, that is, investigating the relation between the occurrence time (age) of chemical reactions and their position in the network (downstream vs. upstream) to draw conclusions about one of the evolution scenarios being at work. The four snapshots in Figure 7 showing the metabolic network in different stages are aligned to a union graph over all generations [42]. Thus, we can see that in the first steps the reactions upward in the network are added. The pathways are formed further in this forward direction. Looking at the last generation, we see that basically all pathways from source to sink follow the forward evolution scenario.

This observation is further supported by the lifetime diagram for all chemical reactions in Figure 8. The reactions are here ordered according to their position in the graph. There is a clear trend of older reactions being on the top (upstream) and younger ones following downstream. The (colored) bar next to the lifetime diagram shows the pattern of the relation between age and position of reactions and metabolites for our example simulation run. The other three bars show the patterns for backward and forward evolution and the patchwork model, respectively. The forward evolution pattern comes closest to the simulated pattern. This illustrates again the speculation from the general analysis that in the early phase of metabolic evolution, forward evolution seems to be dominant. However, for metabolites we do not see a clear relation between the position along pathways in the network and their first appearance in the system. As in the general results, a much more mixed picture is observed for the metabolites. Therefore, no clear explanation can be made for the metabolite constitution.

Another, more complex setting is used in a simulation run in which we investigate the evolutionary history of the involved genes/enzymes, depicted in the catalytic function genealogy for all generations (Figure 9). The simulation takes the same five input molecules from the above general

a



b

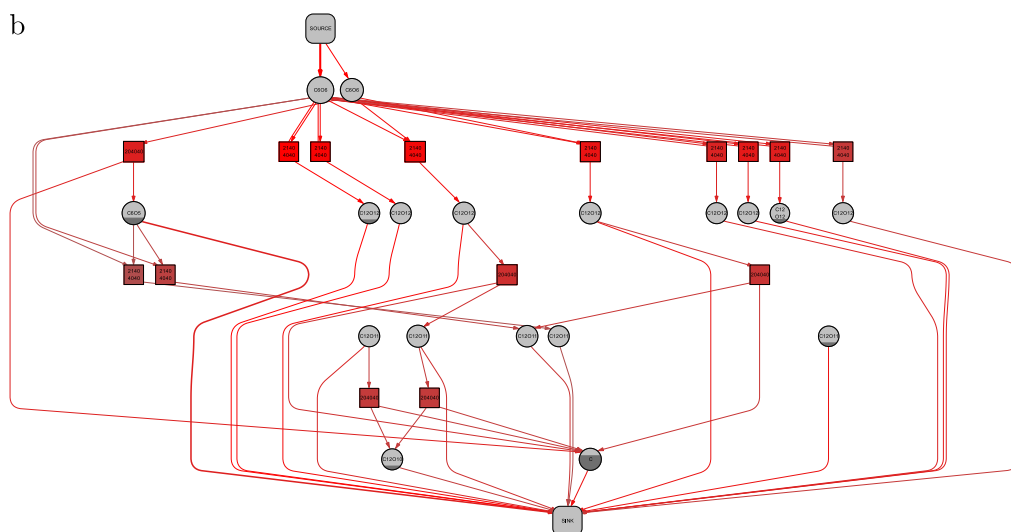
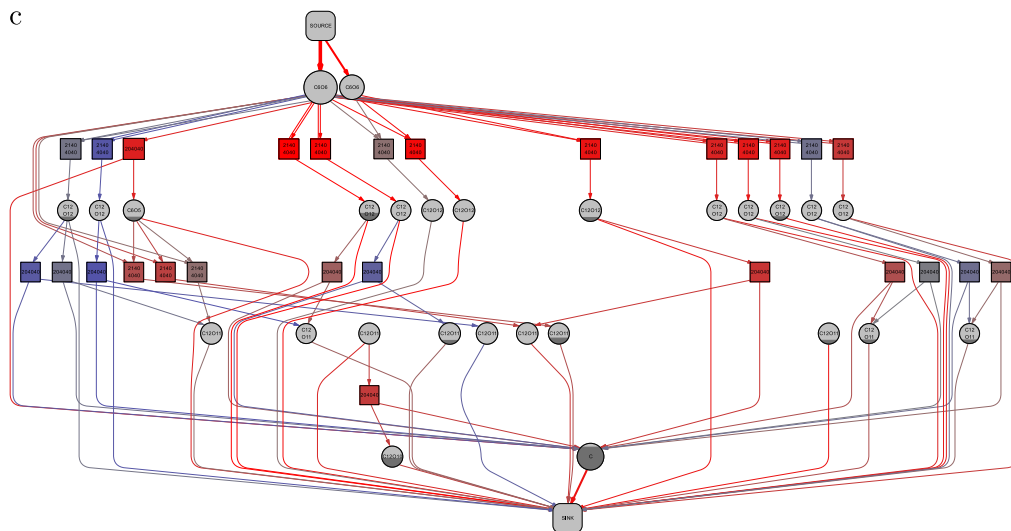


Figure 7. A series of simulated metabolic networks after (a) 10 ($m = 11$, $r = 8$), (b) 30 ($m = 18$, $r = 15$), (c) 66 ($m = 27$, $r = 27$), and (d) 100 generations with a size ($m = 30$, $r = 30$) of 30 metabolites and 30 reactions. Squares represent chemical reactions; circles represent metabolites. Metabolites involved in a reaction are connected to it in the network graph. The size of the nodes and the width of the edges encode for the number of extreme pathways in which the respective object is involved. In the electronic version, the coloring for the reactions encodes their age, where red stands for older (occurrence in early generation) and blue for newer (later generation) reactions.

study, but with a higher mutation and duplication rate, and runs for a total of 2000 generations. Our simulation framework allows us to study the emergence of divergence and convergence of catalytic functions [1], since we can record the genealogy of each gene (reaction catalyst) throughout a simulation run, and we can utilize the ITS classification of the catalyzed reaction as a representation of the enzymatic function. Divergence of a function is caused by gene duplication followed by sequence mutations, creating functionally different but structurally related catalysts. Convergence of a function occurs when catalysts from genealogically unrelated genes independently accumulate mutations resulting in the catalysis of the same reaction (or class of reactions). In Figure 9 convergence events are marked by circles. A small selection of divergence events, which are very frequent in our simulations, are marked by broken circles.

c



d

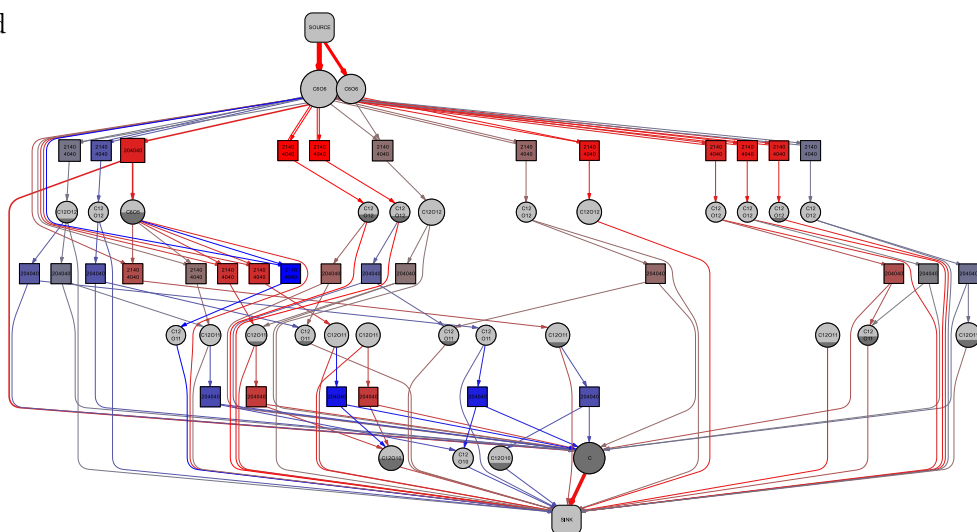


Figure 7. (continued).

In Figure 10a, the history for one specific gene/enzyme (ITS code: 404040) is shown, with the ITS structures of the enzymes leading to the formation of this enzyme and those that originated from it. The divergence of the enzyme is always preceded by an increase in the number of genes, either through duplication or through convergence of other genes/enzymes. Furthermore, the analysis of the functional transitions on the basis of the ITS graphs reveals that catalysts can alter their substrate specificity by small changes of the context of the graph rewrite rule, which is the necessary precondition for the applicability of the graph transformation rule. In our example, most of the adjacent enzymes have similar ITS structures and show agreement in most parts of the substrate structure as well as in the reaction mechanism. However, the first and the last adjacent enzyme do not show any significant similarity to the ITS structure of the studied enzyme. Thus, we will only discuss the other four transitions.

The first of these transitions is described in Figure 10b, with the ITS codes, the ITS structures, and the corresponding reaction mechanisms of the enzymes, as well as a sample reaction using one of the five food metabolites catalyzed by them. Here, only one atom in the context is changed (from

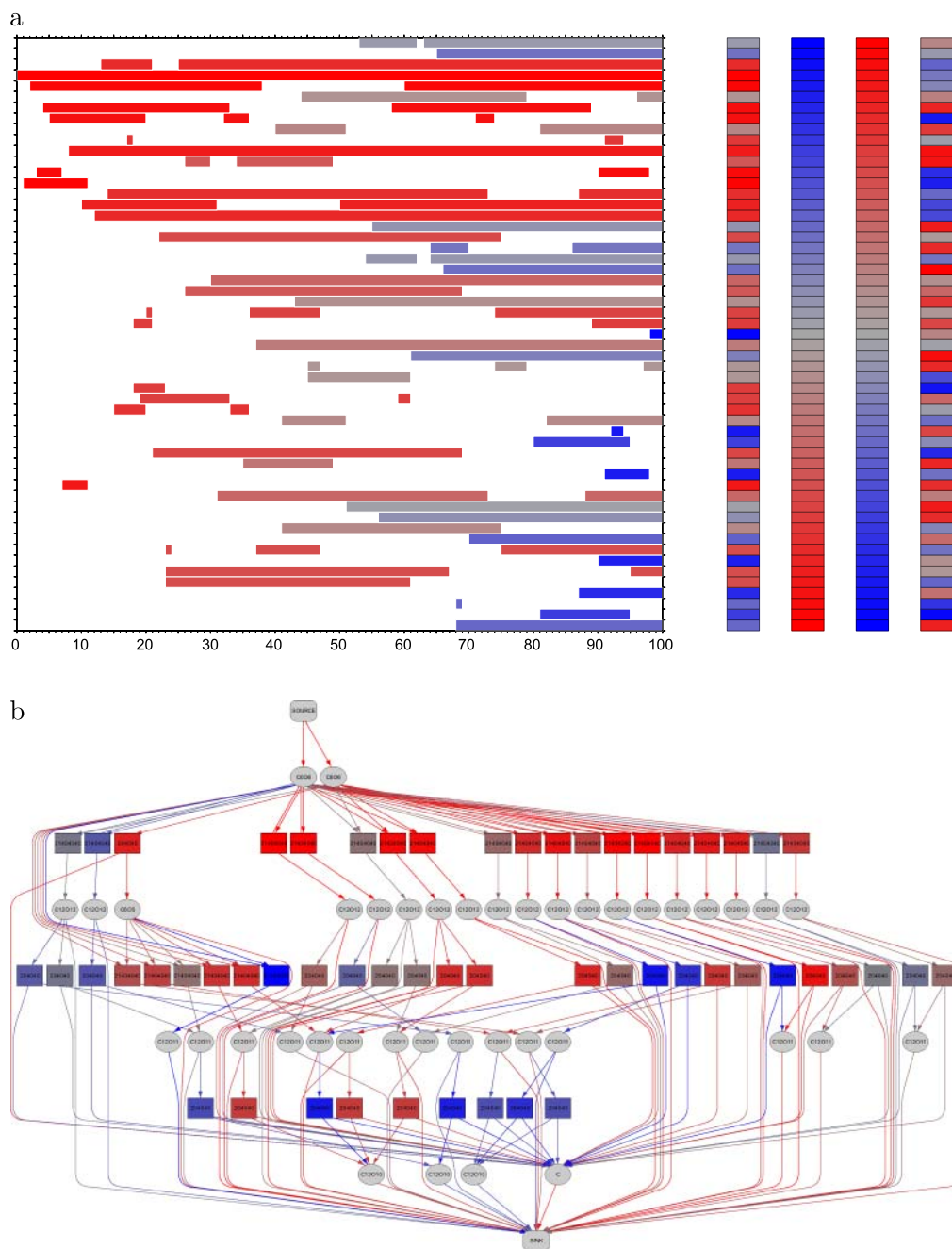


Figure 8. Lifetime diagrams for reactions and metabolites: (a) lifetime of reactions, (b) union network graph over all 100 generations, (c) lifetime of metabolites. The reactions and metabolites (rows) in the diagrams are positioned corresponding to their position in the union network graph, that is, reactions/metabolites close to the source metabolites are in upper positions, and reactions/metabolites close to the sink metabolites are placed at the bottom. The rows have entries if the corresponding reaction/metabolite was present at a certain generation (columns 1–100). For the electronic version, we use the same coloring scheme as above: Older reactions/metabolites are red; newer, blue. The colored bars show the age distribution of reactions in the network in the same order as in the lifetime overview. The first bar represents our results, following the pattern for backward evolution, forward evolution, and the patchwork model.

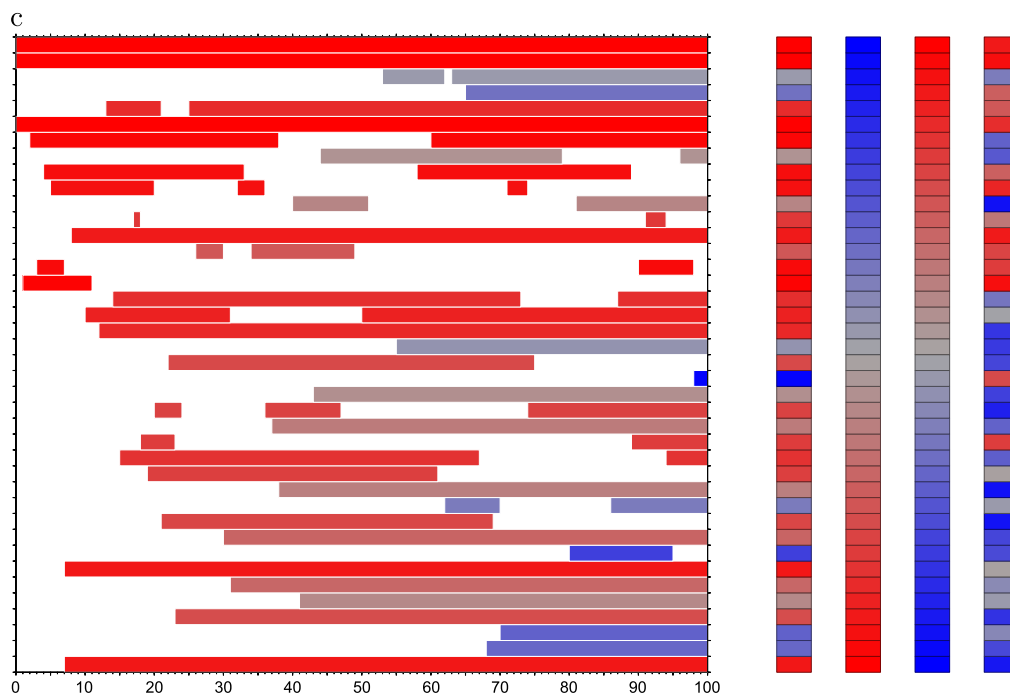


Figure 8. (continued).

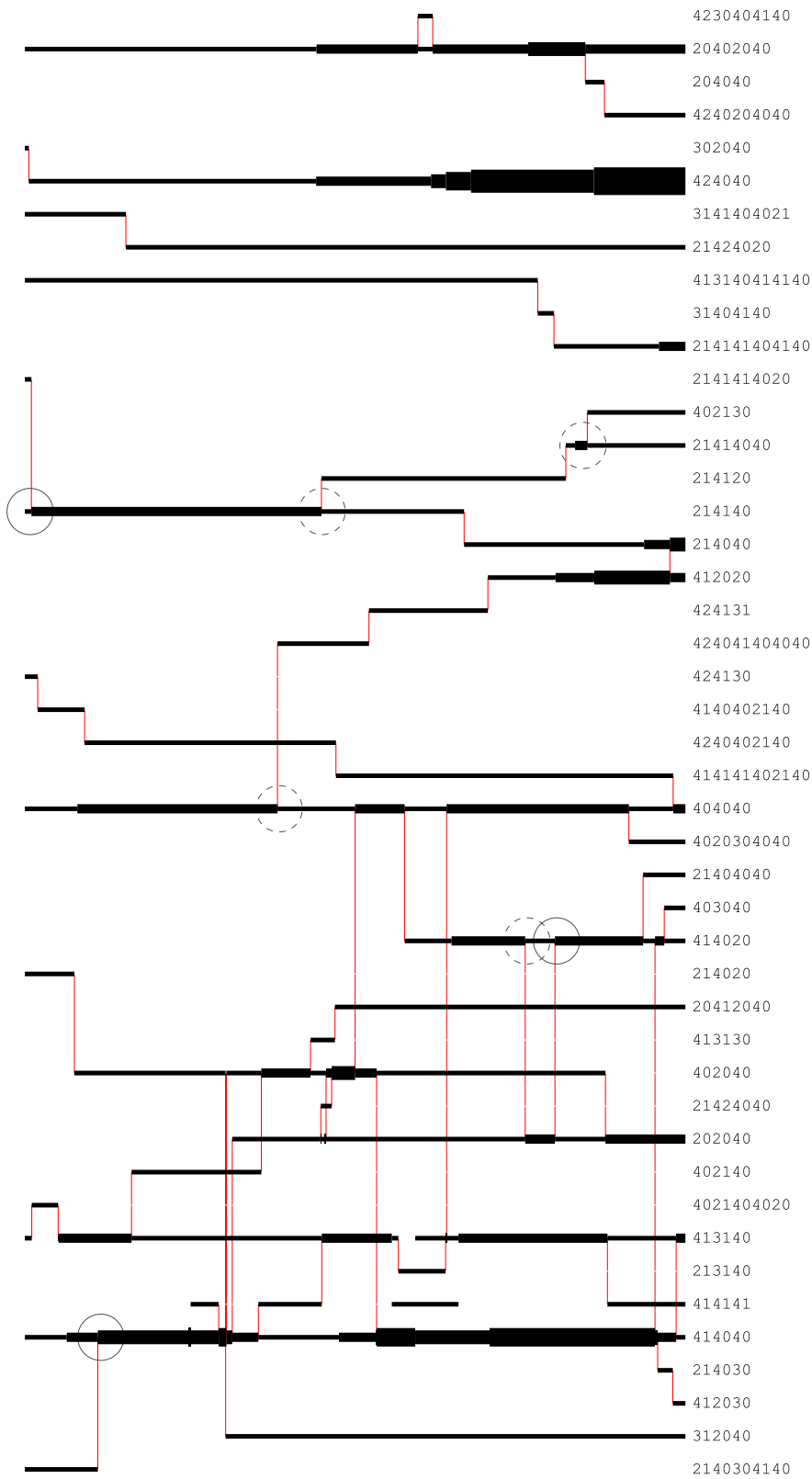
O to C) while all other parts remain the same. This preserved the exact same reaction mechanism. Nevertheless, both enzymes react with different metabolites from the original set of food molecules. The enzyme with the ITS code 402040 reacts with phthalic anhydride, while 404040 is able to make use of methylbutadiene (see example in (b) and (c)) and cyclohexa-1,3-diene (see example in (d) and (e)).

In the next transition (Figure 10c) one additional bond is introduced into the context of the reaction, increasing the substrate specificity of the new enzyme and also resulting in a significant change in the product structure, despite keeping the overall reaction mechanism. The new enzyme uses ethenol, which is also available among the five food metabolites.

The transition in Figure 10d causes a loss in substrate specificity through removal of two bonds from the reaction context. As before, the change of the product structure is more significant than the comparatively moderate change in the substrate molecule. However, this enzyme is not able to use any of the original food metabolites.

A more interesting case is the last of the four transitions (Figure 10e). Although the ITS structures of the enzymes before and after the transition seem very different, a large part of the reaction mechanism is retained and the change in the substrate binding can be described as an increase of substrate specificity by adding two atoms (N and C) to the context of the reaction. The upper parts of the substrate, product, and reaction are the same in both enzymes; only the lower part of the new enzyme differs from that of the original enzyme. The mutated enzyme is able to make use of the same food metabolite (phthalic anhydride) as the enzyme in (b) and yields similar product metabolites, but has a more restrictive context.

These examples demonstrate that, in our simulation universe, relatively small changes in the gene sequence can lead to new enzymes with typically similar reaction mechanisms but different substrate specificity. This in turn sometimes causes quite drastic changes in the reactions that are actually catalyzed. The influence of enzyme promiscuity on enzyme evolution, enzyme engineering, and biocatalysis [3, 7, 28, 32, 36] has been discussed widely throughout the literature for quite a long time. For instance, many enzymes exhibit so called substrate promiscuity, that is, they perform the same



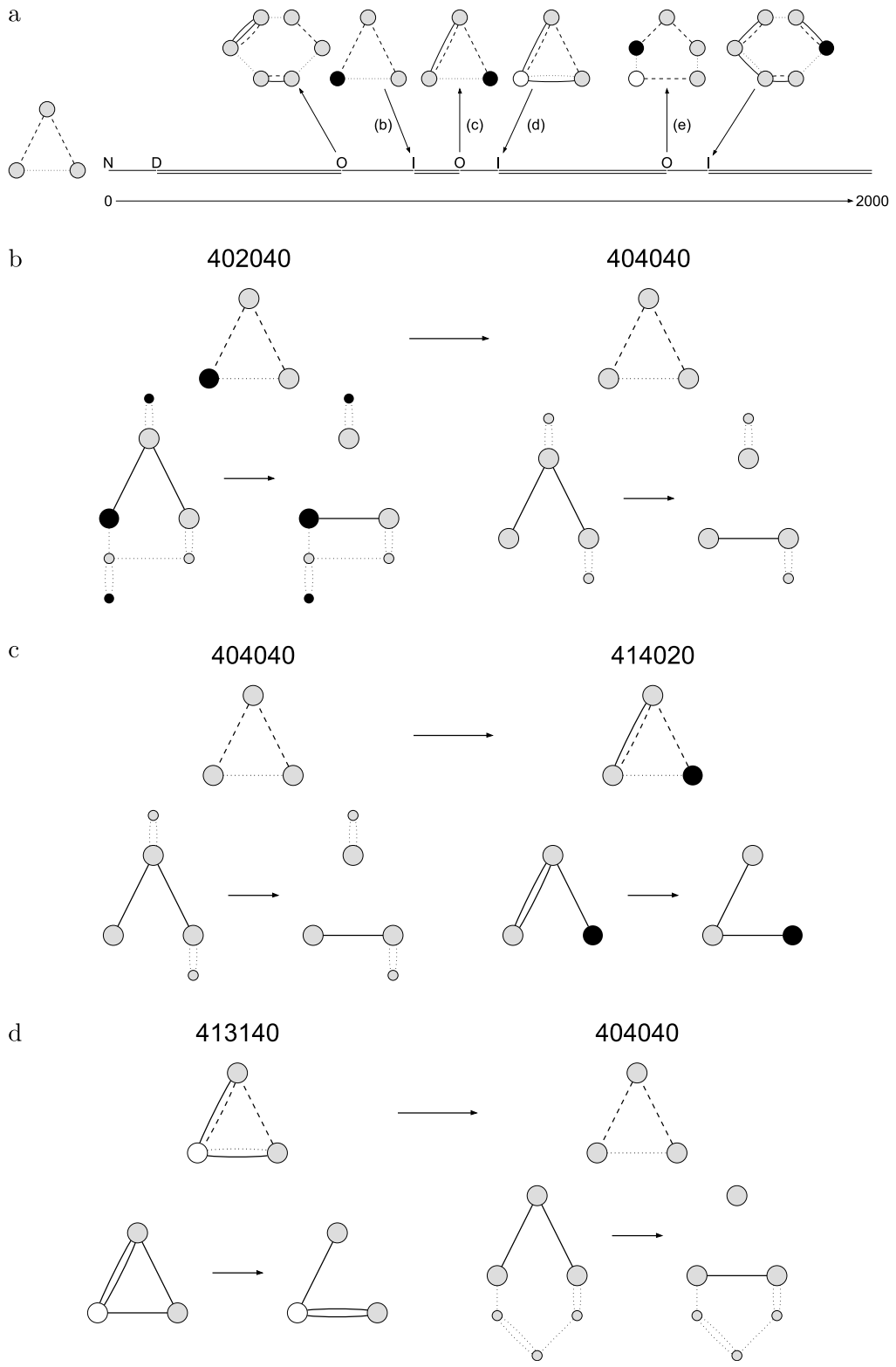
chemical transformation on a wide range of substrates. An impressive natural example for this type of enzyme promiscuity is methane monooxygenase (EC 1.14.13.25), which hydroxylates approximately 150 different alkane substrates in addition to its major substrate methane. This characteristic of natural enzyme function is fully represented in our model and is as well a target of evolutionary change. Other forms of enzyme promiscuity include the occurrence of two different reaction mechanisms. In that case, either the two mechanisms are implemented by the same residues in the active site (thymine hydroxylase EC 1.14.11.6), or the residues in the active site are used in different mechanistic contexts. Such forms of promiscuity do not have a representation within our model. Notably, the sample enzyme and its six neighboring enzymes (connected through functional transitions in this sample simulation) are already able to catalyze chemical reactions using four of the five originally added food metabolites. While most functional transitions from one enzyme to another introduce only a little innovation into the reaction repertoire of the metabolic system, some give access to previously unreachable parts of the existing chemistry.

4 Conclusions

We have introduced a simulation tool that models the early evolution of metabolism in a quite realistic setting and provides many tools for the detailed investigation of metabolic evolution. The evaluation of the simulation data required support by a visualization framework that facilitates the exploration of the data from different points of view and with different levels of detail. The evolution of a network is expressed in the dynamics of a changing graph by the node connectivity and the attribute values associated with nodes and edges, respectively. In our visual analysis approach, we followed Ben Shneiderman's mantra of information visualization: Overview first, zoom and filter, details on demand [47]. The visual exploration of the simulation data thus follows a top-down approach, which has been proven useful when exploring multilevel and multidimensional data. Summarizing time-dependent statistics and lifetime diagrams (e.g., see Figure 8) depicting the temporary presence of nodes in the graph helped to gain insight into the topological dynamics of the evolving metabolic network and eventually allowed the visual identification of "interesting" points in time. For these selected situations, evolutionary transitions in pathways can be analyzed in depth, down to the detailed level of exploring the structure and history of individual metabolic pathways and clusters of pathways involving a particular metabolite. We learned that interactivity plays a crucial role in the exploratory phase of the data analysis, due to the multiscale nature of our simulation framework. It was successfully implemented using the linked view method for intuitive navigation in time as well as within a selected network configuration [42]. We can conclude that efficient visualization is an indispensable tool for exploration and hypothesis formulation in complex simulations of molecular processes in general; see, for example [20, 24, 30, 42], for related applications.

Using both simple examples and a series of more complex simulation runs, the evolution of the components on the small scale (metabolites, enzymes) as well as on systems (pathways, networks) was investigated. The analysis of the genes' histories showed all different kinds of evolutionary events, such as convergence, duplication, and divergence, and many different functional transitions from one gene/enzyme to another, increasing the substrate specificity or changing the reaction chemistry. The simulations further allow us to discriminate between different scenarios for the evolution of metabolic

Figure 9. Genealogy of catalytic functions and gene dosage over 2000 generations. Each row represents an observed catalytic function. Black horizontal lines indicate time intervals in which genes coding for that catalytic function were present in the genome (0–200, from left to right). The thickness of the black lines indicates the number genes with a given function. Thin vertical (red) lines indicate points where the accumulation of mutations caused a transition between catalytic functions. If the number of a gene's copies in a function class increases without a transition from another gene, then the increase is due to a gene duplication. A new gene can be created in the genome through the fortuitous formation of a TATA box. Conversely, a gene can vanish if its TATA box is destroyed by mutation. On the left of the chart a numerical encoding of the graph transformations performed by the enzyme is plotted.



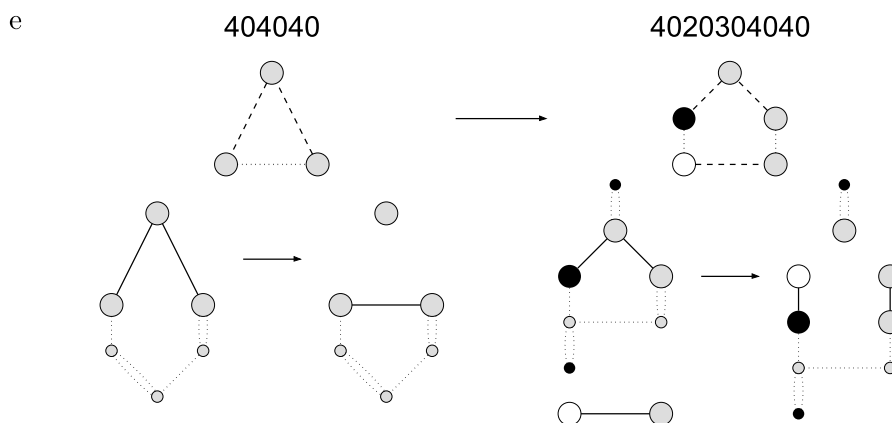


Figure 10. (continued).

pathways. Based on the observations from this study, we argue that the different evolutionary hypotheses can be reconciled, in that they act in different phases of evolution, so that in different scenarios we might observe another strategy at work. Here, we suggest that forward evolution dominates in the earliest steps and is then superseded by a phase of enzyme recruitment—however, leaving behind a trace in the form of a core set of forward-evolved pathways. We also find that the depletion of important food metabolites introduces backward-evolved pathways. However, the formation is driven by enzyme recruitment rather than being a formation from scratch according to the retrograde hypothesis.

To further test these and other hypotheses on the evolution and formation of metabolic pathways, we intend to simulate a number of different scenarios with changing parameters (mutation rate, duplication rate), define other goals for the organisms (production of one specific metabolite, biomass, or energy), and increase the complexity of the simulation runs (length and number of input molecules).

Albeit our simulation environment is still a drastic simplification of chemistry, it is realistic enough to investigate the evolution of early metabolism. Detailed computer simulations, such as the ones reported here, are likely to provide new insights into the general evolutionary mechanisms governing biological systems, in particular in regimes that are not readily observable. Our approach of a realistic, yet computationally feasible, model appears to be a promising step in this direction.

5 Additional Files

An animated movie of an example network evolution simulation can be found at <http://www.bioinf.uni-leipzig.de/~alexander/animation.avi>.

Figure 10. (a) Evolutionary history of a particular enzyme (ITS: 404040). We show the ITS structure of this enzyme and all adjacent enzymes, which arose in a long (2000 generations) simulation run. The ITS structure is depicted as a graph. Lines: solid, reaction context; dashed, bonds that are broken by reaction; dotted, bonds that are created. Circles: black, oxygen; gray, carbon; white, nitrogen. Evolutionary events are marked in the timeline when they occur: N, new occurrence; D, duplication, I (for “in”), convergent event, O (for “out”), divergent event. The number of lines parallel to the timeline indicates the number of gene copies for that enzyme. Four of the six functional transitions are depicted in panels (b)–(e) with the ITS codes (top), the ITS structures (middle), and the reaction mechanisms (bottom) of the two adjacent enzymes. The actual reaction mechanism is represented by the big circles and solid lines only. The small circles and dotted lines give a sample reaction using one of the original food metabolites. In (b) the substrate changes in only one atom position (O to C), in (c) the substrate specificity increases through addition of a bond to the context, in (d) the substrate specificity decreases through removal of two bonds from the context, and in (e) the substrate specificity increases through addition of two atoms (N and C) to the context.

Acknowledgments

This work has been funded by the Volkswagen Stiftung under grant I/82 719 and by the Vienna Science and Technology Fund (WWTF) MA07-30, and the COST-Action CM0703 “Systems Chemistry.”

References

1. Almonacid, D. E., Yera, E. R., Mitchell, J. B., & Babbitt, P. C. (2010). Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: Implications for classification of enzyme function. *PLoS Computational Biology*, 6(3), e1000700.
2. Benkő, G., Flamm, C., & Stadler, P. F. (2003). A graph-based toy model of chemistry. *Journal of Chemical Information and Computer Science*, 43, 1085–1093.
3. Bornscheuer, U. T., & Kazlauskas, R. J. (2004). Catalytic promiscuity in biocatalysis: Using old enzymes to form new bonds and follow new pathways. *Angewandte Chemie (Int. Ed. Engl.)*, 43(45), 6032–6040.
4. Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., & Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. *International Journal Biochemistry & Cell Biology*, 41, 285–297.
5. Ciliberti, S., Martin, O. C., & Wagner, A. (2007). Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences of the U.S.A.*, 104, 13591–13596.
6. Copley, R. R., & Bork, P. (2000). Homology among ($\beta\alpha$)₈-barrels: Implications for the evolution of metabolic pathways. *Journal of Molecular Biology*, 303, 627–641.
7. Copley, S. D. (2003). Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinions in Chemical Biology*, 7(2), 265–272.
8. Cordon, F. (1990). *Tratado evolucionista de biología*. Madrid: Aguilar Ediciones.
9. Fani, R., & Fondi, M. (2009). Origin and evolution of metabolic pathways. *Physics of Life Reviews*, 6, 23–52.
10. Faulon, J.-L., & Sault, A. G. (2001). Stochastic generator of chemical structure. 3. Reaction network generation. *Journal of Chemical Information and Computer Science*, 41, 894–908.
11. Feist, A. M., & Palsson, B. Ø. (2010). The biomass objective function. *Current Opinions in Microbiology*, 13, 344–349.
12. Flamm, C., Ullrich, A., Ekker, H., Mann, M., Högerl, D., Rohrschneider, M., Sauer, S., Scheuermann, G., Klemm, K., Hofacker, I. L., & Stadler, P. F. (2010). Evolution of metabolic networks: A computational framework. *Journal of Systems Chemistry*, 1, 4.
13. Fontana, W., Schnabl, W., & Schuster, P. (1989). Physical aspects of evolutionary optimization and adaptation. *Physical Review A*, 40, 3301–3321.
14. Fontana, W., & Schuster, P. (1998). Continuity in evolution: On the nature of transitions. *Science*, 280, 1451–1455.
15. Fothergill-Gilmore, L. A., & Michels, P. A. (1993). Evolution of glycolysis. *Progress in Biophysics and Molecular Biology*, 59(2), 105–235.
16. Fujita, S. (1986). Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *Journal of Chemical Information and Computer Science*, 26, 205–212.
17. Gagneur, J., & Klamt, S. (2004). Computation of elementary modes: A unifying framework and the new binary approach. *BMC Bioinformatics*, 5.
18. Gerlt, J. A., & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Reviews in Biochemistry*, 70, 209–246.
19. Granick, S. (1957). Speculations on the origins and evolution of photosynthesis. *Annals of the New York Academy of Sciences*, 69, 292–308.
20. Heine, C., Scheuermann, G., Flamm, C., Hofacker, I. L., & Stadler, P. F. (2006). Visualization of barrier tree sequences. *IEEE Transactions on Visualization and Computer Graphics*, 12(5), 781–788.
21. Hendrickson, J. B. (1997). Comprehensive system for classification and nomenclature of organic reactions. *Journal of Chemical Information and Computer Science*, 37, 852–860.

22. Hendrickson, J. B., & Miller, T. M. (1990). Reaction indexing for reaction databases. *Journal of Chemical Information and Computer Science*, 30, 403–408.
23. Herges, R. (1994). Coarctate transition states: The discovery of a reaction principle. *Journal of Chemical Information and Computer Science*, 34, 91–102.
24. Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., & Stadler, P. F. (2010). BarMap: RNA folding on dynamic energy landscapes. *RNA*, 16(7), 1308–1316.
25. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Chemical Monthly*, 125, 167–188.
26. Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences of the U.S.A.*, 31, 153–157.
27. Hrmova, M., De Gori, R., Smith, B. J., Fairweather, J. K., Driguez, H., Varghese, J. N., & Fincher, G. B. (2002). Structural basis for broad substrate specificity in higher plant beta-D-glucan glucohydrolases. *Plant Cell*, 14, 1033–1052.
28. Hult, K., & Berglund, P. (2007). Enzyme promiscuity: Mechanism and applications. *Trends in Biotechnology*, 25(5), 231–238.
29. Huynen, M. A., Stadler, P. F., & Fontana, W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation. *Proceedings of the National Academy of Sciences of the U.S.A.*, 93, 397–401.
30. Jänicke, S., Heine, C., Hellmuth, M., Stadler, P. F., & Scheuermann, G. (2010). Visualization of graph products. *IEEE Transactions on Visualization and Computer Graphics (IEEE Information Visualization Conference)*, 16(6), 1082–1089.
31. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology*, 30, 409–425.
32. Khersonsky, O., Roodveldt, C., & Tawfik, D. S. (2006). Enzyme promiscuity: Evolutionary and mechanistic aspects. *Current Opinions in Chemical Biology*, 10(5), 498–508.
33. Khersonsky, O., & Tawfik, D. S. (2010). Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Reviews in Biochemistry*, 79, 471–505.
34. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K., & Copley, S. D. (2010). Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis. *Molecular Systems Biology*, 6, 436.
35. Morowitz, H. J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. *Complexity*, 4, 39–53.
36. O'Brien, P. J., & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chemical Biology*, 6(4), R91–R105.
37. Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, 28, 245–248.
38. Ourisson, G., & Nakatani, Y. (1994). The terpenoid theory of the origin of cellular life: The evolution of terpenoids to cholesterol. *Chemical Biology*, 1(1), 11–23.
39. Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D., & Kozarich, J. W. (1993). On the origin of enzymatic species. *Trends in Biochemical Sciences*, 18(10), 372–376.
40. Pfeiffer, T., Soyer, O. S., & Bonhoeffer, S. (2005). The evolution of connectivity in metabolic networks. *PLoS Biology*, 3, e228.
41. Reidys, C. M., & Stadler, P. F. (2002). Combinatorial landscapes. *SIAM Review*, 44, 3–54.
42. Rohrschneider, M., Ullrich, A., Kerren, A., Stadler, P. F., & Scheuermann, G. (2010). Visual network analysis of dynamic metabolic pathways. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammoud, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, & L. Avila (Eds.), *Advances in Visual Computing (ISVC 2010)* (pp. 316–327). Berlin: Springer.
43. Samal, A., Rodrigues, J. F. M., Jost, J., Martin, O. C., & Wagner, A. (2010). Genotype networks in metabolic reaction spaces. *BMC Systems Biology*, 4, 30.
44. Schmidt, S., Sunyaev, S., Bork, P., & Dandekar, T. (2003). Metabolites: A helping hand for pathway evolution? *Trends in Biochemical Sciences*, 28, 336–341.

45. Schuster, P. (1999). Chance and necessity in evolution: Lessons from RNA. *Physica D*, 133, 427–452.
46. Schuster, P., Fontana, W., Stadler, P. F., & Hofacker, I. L. (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings of the Royal Society B: Biological Sciences*, 255, 279–284.
47. Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)* (pp. 336–343).
48. Stadler, B. M. R., Stadler, P. F., Wagner, G., & Fontana, W. (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213, 241–274.
49. Stadler, P. F. (1999). Fitness landscapes arising from the sequence-structure maps of biopolymers. *Journal of Molecular Structure*, 463, 7–19.
50. Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L., & Schuster, P. (1996). Algorithm independent properties of RNA structure prediction. *European Biophysics Journal*, 25, 115–130.
51. Ullrich, A., & Flamm, C. (2008). Functional evolution of ribozyme-catalyzed metabolisms in a graph-based toy-universe. In S. Istrail (Ed.), *Proceedings of the 6th International Conference on Computational Methods in Systems Biology (C SMB)* (pp. 28–43). Berlin: Springer.
52. Ullrich, A., & Flamm, C. (2010). A sequence-to-function map for ribozyme-catalyzed metabolisms. In *Proceedings of ECAL 2009*. Berlin: Springer.
53. Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., & Olsson, M. H. (2006). Electrostatic basis for enzyme catalysis. *Chemical Reviews*, 106, 3210–3235.
54. Weberndorfer, G., Hofacker, I. L., & Stadler, P. F. (2003). On the evolution of primitive genetic codes. *Origins of Life and Evolution of the Biosphere*, 33, 491–514. SFI preprint #02-08-034.
55. Ycas, M. (1974). On earlier states of the biochemical system. *Journal of Theoretical Biology*, 44, 145–160.